
Constraining the Phrase-Based Joint Probability Model

Alexandra Birch, Chris Callison-Burch, and Miles Osborne

August 10, 2006



Overview

- Motivation:
 - Current phrase-based SMT models use ad-hoc heuristics
 - Marcu and Wong's joint phrase model is simpler and theoretically preferable
- Problem: Computational complexity of estimating phrase-to-phrase probs
- Our solution: Constrain joint phrase model using external knowledge to limit what phrases can be aligned
- Experimental results and conclusions

Current Phrase-Based Models

- The standard method for training a phrase-based model is
 1. Use Giza++ to train IBM Models in the $e \rightarrow f$ direction
 2. Use Giza++ to train IBM Models in the $f \rightarrow e$ direction
 3. Output single best $e \rightarrow f$ and $f \rightarrow e$ word-alignment for each sentence pair
 4. Use heuristics to combine the $e \rightarrow f$ and $f \rightarrow e$ alignments
 5. Use phrase extraction method to list all phrase pairs
 6. Calculate MLE probabilities by counting phrase pair co-occurrences

Problems with Standard Approach

- We the overcome deficiencies in IBM models with *ad hoc* heuristics
- We commit to a *single* alignment for each sentence, disregarding others
- Our phrase extraction methods are *arbitrarily* defined

We're doing statistical machine translation but not estimating our probabilities in a very rigorous manner.

Alternative: Estimate phrase probabilities properly

- Currently use cludgy methods for extracting phrase alignments and probabilities from word-alignments
- Instead try Marcu and Wong's (2002) phrase-based, joint probability model for statistical machine translation
- Estimates phrase alignments and phrase translation probabilities directly

Joint Model

$$p(F, E) = \sum_{C \in \mathcal{C}} \prod_{\langle \bar{e}_j, \bar{f}_j \rangle \in C} p(\langle \bar{e}_j, \bar{f}_j \rangle)$$

- Variables:
 - F is a foreign sentence
 - E is an English sentence
 - \bar{e} is a sequence of words in E
 - \bar{f} is a sequence of words in F
 - C is a set of $\langle \bar{e}_j, \bar{f}_j \rangle$ which cover all words in E, F
 - \mathcal{C} is all such sets
- Use EM to estimate $p(\langle \bar{e}_j, \bar{f}_j \rangle)$ for all phrases in our corpus

Advantages of the Joint Model

- vs. IBM models
 - Allows phrase-to-phrase alignments
 - Eliminates need for strange parameters like fertility, spurious word probability
 - Reduces dependency on distortion
- vs. Standard phrase-base approach
 - Don't need to perform ad hoc symmetrization
 - Don't need to define arbitrary phrase extraction from single word alignment
- Estimate phrase translation probabilities directly from corpus

Problems with the Joint Model

- Searches the entire phrasal alignment space
- Complexity explodes - all possible segmentations and their alignments

Problems with the Joint Model

- Searches the entire phrasal alignment space
- Complexity explodes - all possible segmentations and their alignments

E Length	F Length	No. Alignments
5	5	6721
10	10	$8.182 * 10^{11}$
20	20	$4.414 * 10^{32}$
40	40	$2.734 * 10^{83}$

Problems with the Joint Model

- Searches the entire phrasal alignment space
- Complexity explodes - all possible segmentations and their alignments

E Length	F Length	No. Alignments
5	5	6721
10	10	$8.182 * 10^{11}$
20	20	$4.414 * 10^{32}$
40	40	$2.734 * 10^{83}$

Number of milliseconds until the sun becomes a red giant and engulfs the Earth
 $\approx 1.5768 * 10^{20}$.

Approximate Training

- In EM we should collect fraction counts over all possible alignments
- Cannot do that, so training can only be approximate
- Marcu and Wong use the following approximations:
 - Limit phrases considered to high frequency phrases
 - Collect fractional counts only over small subset of alignments
 - Limited training sentences to max of 20 words

Approximate training also has to be used in IBM Models

Problems Remain

- Still have difficulty scaling the method due to time and memory requirements
- Less likely to find optimal alignments

Solution: Constraints

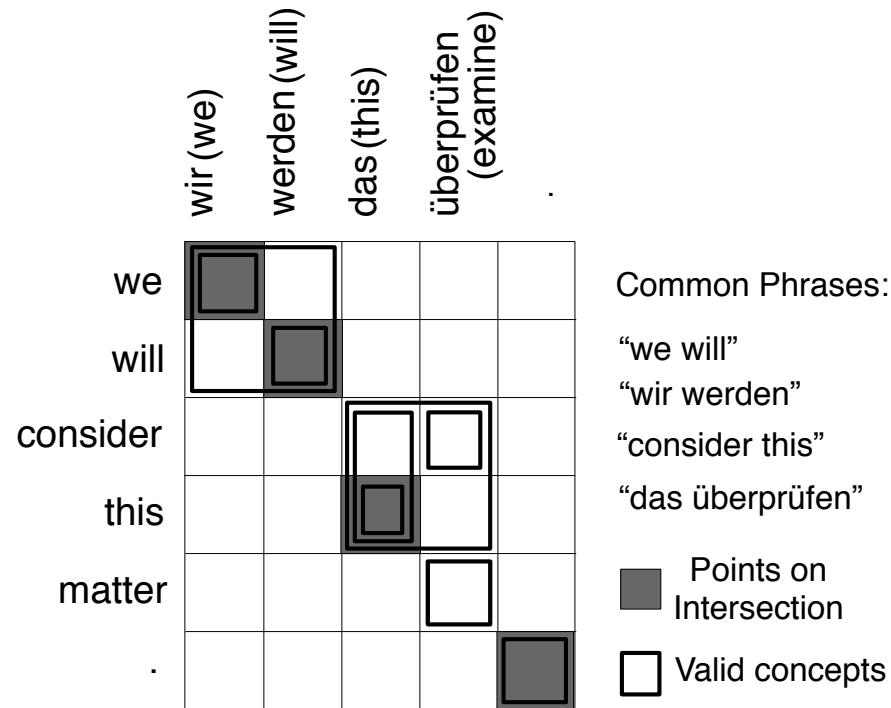
- We propose to constrain what alignments are considered
- In addition to looking only at high frequency phrases, we fix certain points in grid, which correspond to high confidence word translations
- Consider only alignments which are consistent with these points

Solution: Constraints

- We propose to constrain what alignments are considered
- In addition to looking only at high frequency phrases, we fix certain points in grid, which correspond to high confidence word translations
- Consider only alignments which are consistent with these points

By consistent we mean that for a phrase alignment $\langle \bar{e}_j, \bar{f}_j \rangle$ to be valid, if any word in \bar{e}_j is part of a fixed point, then its corresponding foreign word must be in \bar{f}_j and vice versa.

Search Space



All valid concepts for constrained joint model

Where Constraints Come From

- Use external knowledge to get these points
 - Bilingual dictionary
 - Identical words
 - Intersection of IBM Models

Effects of the Constraints

- Restricting search space to areas with most probability mass
- Improve parameters
- Scale model to larger corpora

Experimental Design

- We compared the unconstrained joint model, our constrained version, and the standard phrase-based models
- Measured Bleu scores for different models, and sizes of the phrase tables
- Trained on small parts of the Europarl German-English parallel corpus
- Performed translation with Pharaoh

Results

Bleu scores

Corpus Size (# sents)	10,000	20,000	40,000
Standard Model	21.69	23.61	25.52
Joint Model			
+ IBM constraints			
+ IBM + Dict + Ident			

Translation table sizes

Corpus Size (# sents)	10,000	20,000	40,000
Standard Model	90k	200k	410k
Joint Model			
+ IBM constraints			
+ IBM + Dict + Ident			

Results

Bleu scores

Corpus Size (# sents)	10,000	20,000	40,000
Standard Model	21.69	23.61	25.52
Joint Model	19.93	-	-
+ IBM constraints			
+ IBM + Dict + Ident			

Translation table sizes

Corpus Size (# sents)	10,000	20,000	40,000
Standard Model	90k	200k	410k
Joint Model	6.17m	-	-
+ IBM constraints			
+ IBM + Dict + Ident			

Results

Bleu scores

Corpus Size (# sents)	10,000	20,000	40,000
Standard Model	21.69	23.61	25.52
Joint Model	19.93	-	-
+ IBM constraints	22.79	24.33	25.99
+ IBM + Dict + Ident			

Translation table sizes

Corpus Size (# sents)	10,000	20,000	40,000
Standard Model	90k	200k	410k
Joint Model	6.17m	-	-
+ IBM constraints	1.45m	2.72m	4.96m
+ IBM + Dict + Ident			

Results

Bleu scores

Corpus Size (# sents)	10,000	20,000	40,000
Standard Model	21.69	23.61	25.52
Joint Model	19.93	-	-
+ IBM constraints	22.79	24.33	25.99
+ IBM + Dict + Ident	23.20	24.96	26.13

Translation table sizes

Corpus Size (# sents)	10,000	20,000	40,000
Standard Model	90k	200k	410k
Joint Model	6.17m	-	-
+ IBM constraints	1.45m	2.72m	4.96m
+ IBM + Dict + Ident	1.09m	2.07m	3.83m

What About Large Corpora?

	BLEU	No. Phrase Pairs
Pharaoh Baseline	26.15	19.04M
Joint Model	25.49	2.28M
Pharaoh +MERT	28.35	19.04M
Joint +MERT	26.17	2.28M

- Whole Spanish-English Europarl training corpus and test set from NAACL 2006 SMT workshop
- Pruned joint model table significantly in order to scale (perhaps too much)
- Strangely, get less gain than standard model when using MERT

Conclusion

- Joint Model has challenges including time complexity, memory
- By applying constraints we've started to overcome them
- Performance of joint model better than the standard models on small data sets
- Joint model is simpler and theoretically preferable
- Joint model is a viable and interesting alternative!

Future Work

- System engineering improvements
- Parallelize, pruning, good initialization and constraints, priors
- Extending to *factored translation models*, using POS sequences to constrain potential alignments

Thank you!